Lorentz Group Equivariant Neural Network for Particle Physics

Alexander Bogatskiy [1] **Brandon Anderson** [2 3] **Jan T. Offermann** [1] **Marwah Roussi** [1] **David W. Miller** [1 4]
Risi Kondor [2 5 6]

Abstract

We present a neural network architecture that is fully equivariant with respect to transformations under the Lorentz group, a fundamental symmetry of space and time in physics. The architecture is based on the theory of the finite-dimensional representations of the Lorentz group and the equivariant nonlinearity involves the tensor product. For classification tasks in particle physics, we demonstrate that such an equivariant architecture leads to drastically simpler models that have relatively few learnable parameters and are much more physically interpretable than leading approaches that use CNNs and point cloud approaches. The competitive performance of the network is demonstrated on a public classification dataset (Kasieczka et al., 2019) for tagging top quark decays given energy-momenta of jet constituents produced in proton-proton collisions.

1. Introduction

The success of CNNs as a method of computer vision has made clear the benefits of explicitly translationally equivariant neural network architectures: there are far fewer learnable parameters and these parameters are organized into much more interpretable structures. The ability to interpret convolutional kernels as images boosted our understanding of why and how such networks operate (Zeiler & Fergus, 2014).

However, there are many relevant problems that exhibit much more complex symmetries than flat images. Such problems may require or benefit from latent space representations that are intimately connected with the theory of the specific underlying symmetry group. Indeed, these symmetries are manifest in the data itself, as each data point is generated by a symmetric process or model. Following this approach, elegant architectures can be advised based on fundamental principles, and the "building blocks" of such architectures are greatly restricted by the imposed symmetries. This is a highly sought-after property in neural network design since it may improve generality, interpretability, and uncertainty quantification, while simplifying the model.

These general ideas have already led to the development of multiple equivariant architectures for sets (permutation invariance) (Zaheer et al., 2017), graphs (graph isomorphisms), 3D data (spatial rotations) (Monti et al., 2017), and homogeneous spaces of Lie groups such as the two-dimensional sphere (Cohen et al., 2018). For more discussion and references see Section 2.

Symmetries play a central role in any area of physics (Frankel, 2004), and as such physics provides the widest variety of symmetry groups relevant in computational problems. In particular, high energy and particle physics involve symmetry groups ranging from U(1), SU(2) and SU(3) to the Lorentz group $\mathrm{SO}(1,3)$, and even more exotic ones like E_8. Architectures that respect these symmetries can provide more sensible and tractable models, whose parameters may be directly interpreted in the context of known physical models, as in the case of CNNs.

Harmonic analysis provides two parallel but theoretically equivalent implementations of group equivariance in neural networks. The first is a natural generalization of CNNs to arbitrary Lie groups and their homogeneous spaces (Cohen & Welling, 2016), where activations are functions on the group, the nonlinearity is applied point-wise, and the convolution is an integral over the group. The second approach works entirely in the Fourier space (Thomas et al., 2018; Anderson et al., 2019), that is, on the set of irreducible representations of the group. It is the latter approach that we adopt in this work due to its direct applicability to vector inputs.

These approaches are general, but here we present the first specific application of a group equivariant architecture in physics. We focus on a particle physics application where

[1]Department of Physics, University of Chicago, Chicago, IL, U.S.A. [2]Department of Computer Science, University of Chicago, Chicago, IL, U.S.A. [3]Atomwise, San Francisco, CA, U.S.A. [4]Enrico Fermi Institute, Chicago, IL, U.S.A. [5]Department of Statistics, University of Chicago, Chicago, IL, U.S.A. [6]Flatiron Institute, Simons Foundation, New York, NY, U.S.A.. Correspondence to: Alexander Bogatskiy <bogatsky@uchicago.edu>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

the data typically contain the energy-momentum 4-vectors of particles produced in collision events at high energy particle accelerators such as the Large Hadron Collider (LHC) at CERN in Geneva, Switzerland, or by simulation software used to model the collision events. Probing elementary particle collisions at high energies is one of the best approaches to discover new small-scale fundamental phenomena, such as the discovery of Higgs boson at the LHC in 2012 (ATLAS Collaboration, 2012; CMS Collaboration, 2012). There the collisions occur 40 million times per second (40 MHz) between clouds of protons traveling at nearly the speed of light. Within each proton-proton bunch collision an average of $O(30)$ individual pairs of protons collide and produce sprays of outgoing particles that are measured by complex detection systems. These *detectors* – such as the general-purpose ATLAS (ATLAS Collaboration, 2008) and CMS (CMS Collaboration, 2008) detectors – have $O(100M)$ individual sensors that record combinations of positions, trajectories, momenta, and energies of outgoing particles. The data obtained from these detectors must therefore be both filtered and processed by automated on-line systems.

The energy-momentum vector of a particle depends on the inertial frame of the observer, and the transformations between these frames are described by the Lorentz group $\mathrm{O}(1,3)$. In addition to regular spatial rotations it contains the so-called *Lorentz boosts*, which make the Lorentz group non-compact. The architecture presented below avoids computational difficulties associated with the non-compactness of the group by working entirely within its finite-dimensional representations. This choice is not only computationally efficient, but also physically sensible.

2. Related Work

There is a large body of work on equivariance in machine learning. Here we mention a few notable publications most closely related to the methods of our work. Equivariance in neural networks was first examined in applications involving finite groups, such as graph isomorphisms (Bruna et al., 2014; Henaff et al., 2015) and permutations (Zaheer et al., 2017). A general approach to group-convolutional neural networks was proposed in (Cohen & Welling, 2016). Equivariant networks with respect to spacial translations and rotations were developed in (Worrall et al., 2017). For rotational symmetries of the 2-dimensional sphere, the importance of the Fourier space spanned by spherical harmonics was realized in (Cohen et al., 2018; Esteves et al., 2018). In (Weiler et al., 2018) this approach was extended to the entire Euclidean group SE(3). A complete description of equivariant networks for scalar fields on homogeneous spaces of compact Lie groups was given in (Kondor & Trivedi, 2018). It was later generalized to general gauge fields in (Cohen et al., 2019a;b).

The parallel approach, where even the nonlinear operations are performed equivariantly in the Fourier space of SO(3), was independently proposed in (Thomas et al., 2018) and (Kondor, 2018). Successful applications of these ideas in computer vision and chemistry were demonstrated in (Kondor et al., 2018; Anderson et al., 2019). While the use of Lorentz-invariant quantities and Lorentz transformations in networks has been demonstrated in (Butter et al., 2018; Erdmann et al., 2019), our work provides the first equivariant neural network architecture for fundamental physics applications.

3. Theory of the Lorentz group

Lorentz transformations Particles moving in laboratories at velocities approaching the speed of light are described by the theory of special relativity. Its mathematical formulation is based on the postulate that space and time are unified into the 4-dimensional *spacetime*, and the Euclidean dot product of vectors is replaced by the Minkowski, or Lorentzian, metric. In the standard Cartesian basis, this metric has the diagonal form $\mathrm{diag}(1,-1,-1,-1)$:

$$(t,x,y,z)\cdot(t',x',y',z') = tt'-xx'-yy'-zz' = \eta_{\mu\nu}x^\mu x'^\nu$$

(here we set the speed of light equal to one and use the Einstein summation convention for repeated indices). Similarly, the energy and momentum of a particle are combined into the energy-momentum 4-vector whose square is also the mass squared of the particle:

$$(E,p_x,p_y,p_z)^2 = E^2 - p_x^2 - p_y^2 - p_z^2 = m^2.$$

An *inertial frame* in this spacetime is a choice of an orthonormal basis $\{\boldsymbol{e}_0,\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3\}$, i.e. $\boldsymbol{e}_a\cdot\boldsymbol{e}_b = \eta_{ab}$, $a,b = 0,\ldots,3$. The components of the metric are the same in any such frame. The *Lorentz group* is defined as the group of linear isomorphisms Λ^μ_ν of the spacetime that map inertial frames to inertial frames, or equivalently, preserve the metric:

$$\Lambda^\lambda_\mu\eta_{\lambda\rho}\Lambda^\rho_\nu = \eta_{\mu\nu}.$$

This group is denoted by $\mathrm{O}(1,3)$ and consists of 4 connected components distinguished by orientations of space and time.

Often one further requires inertial frames to be positively oriented and positively time-oriented. That is, all orthonormal bases are similarly oriented and the timelike basis vector in each of them ($\boldsymbol{e}_0$) belongs to the future light cone (i.e. its temporal component is positive). Restricting Lorentz transformations to only such frames (which amounts to requiring $\det\Lambda = 1$ and $\Lambda^0_0 > 0$), one obtains the *proper orthochronous Lorentz group* $\mathrm{SO}^+(1,3)$, which is the connected component of the identity in $\mathrm{O}(1,3)$. From here on in this text, this is the group we will call the "Lorentz group". The basic principle of special relativity is that all laws of

physics appear equivalent to observers in all inertial frames. This makes the Lorentz group the fundamental symmetry in relativistic physics.

The group SO(3) of spatial rotations (acting on x, y, z in a chosen inertial frame) is a subgroup of the Lorentz group. In addition to these rotations, it contains the so-called *Lorentz boosts* which transform between inertial frames of observers moving relative to each other at a relative velocity $\boldsymbol{\beta} = \boldsymbol{v}/c$ (in units of the speed of light c). Namely, given two inertial frames $\{\boldsymbol{e}_i\}_{i=0}^3$ and $\{\boldsymbol{e}'_i\}_{i=0}^3$, the relative velocity vector $\boldsymbol{\beta}$ and the *boost factor* γ are defined by $\boldsymbol{e}'_0 = \gamma \boldsymbol{e}_0 + \sum_{i=1}^3 \gamma\beta_i \boldsymbol{e}_i$. Since $\boldsymbol{e}'_0$ has unit norm, the boost factor is related to $\boldsymbol{\beta}$ by $\gamma = (1-\beta^2)^{-1/2}$. Now, if one rotates the spatial axes so that $\boldsymbol{\beta} = (\beta, 0, 0)$ then the Lorentz transformation between these two frames is the matrix

$$\begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In the limit of speeds much lower than the speed of light, $\beta \to 0$, $\gamma \to 1$, and this matrix becomes the identity matrix, returning us to Galilean mechanics. Therefore the appearance of the boost factor γ is a signature of relativistic physics. Lorentz boosts are sometimes called "hyperbolic rotations" because their components can be expressed as $\gamma = \cosh\alpha$ and $\gamma\beta = \sinh\alpha$ in terms of a *rapidity* α. However, note that Lorentz boosts with more than one spatial dimension do not form a subgroup.

Representations Recall that a finite-dimensional representation of a Lie group G is a finite-dimensional vector space V with an action of the group via invertible matrices, that is, a smooth homomorphism $\rho : G \to \mathrm{GL}(V)$ (for some introductions to representations see (Barut & Rączka, 1977; Hall, 2015), or (Diaconis, 1988) for a more applied focus). All activations in our neural network will belong to various finite-dimensional representations of the Lorentz group. Importantly, such representations are *completely reducible*, which means that they are isomorphic to direct sums of irreducible representations (the isomorphism may not be unique). An irreducible representation ("irrep") is one without any invariant subspaces, i.e. subspaces $W \subset V$ such that $\rho(g) \cdot w \in W$ for all $g \in G$ and $w \in W$. Writing the decomposition of V as $V \cong \bigoplus_\alpha R_\alpha^{\oplus\tau_\alpha}$, where R_α's for different α are non-isomorphic irreps of G, we call τ_α the *multiplicity* of R_α in V. Written in terms of the subspaces $V_\alpha \cong R_\alpha^{\oplus\tau_\alpha}$, this decomposition $V = \bigoplus_\alpha V_\alpha$ is called the *isotypic decomposition* of V. Complete reducibility provides a natural basis for storing arbitrary representation vectors, therefore we will now review the classification of finite-dimensional irreps of $\mathrm{SO}^+(1,3)$.

The representation theory of the Lorentz group becomes slightly simpler if we pass to its universal covering group. For SO(3) the universal covering group is SU(2), also known as the 3-dimensional spin group, and for $\mathrm{SO}^+(1,3)$, which is isomorphic to the projective special linear group $\mathrm{PSL}(2,\mathbb{C})$, it is $\mathrm{SL}(2,\mathbb{C})$. Both of these are double covers, i.e. we have $\mathrm{SO}(3) \cong \mathrm{SU}(2)/\{\pm I\}$ and $\mathrm{SO}^+(1,3) \cong \mathrm{SL}(2,\mathbb{C})/\{\pm I\}$. Each irrep of the original group can be extended to an irrep of its double cover, but the double cover generally has more irreps (Gelfand et al., 1963). In physics, the extra "double-valued" irreps obtained by passing to the double cover are called *spinor representations*. Since $\mathrm{SL}(2,\mathbb{C})$ is the complex form of SU(2), the finite-dimensional representations of these two groups are very closely related. These labels of the irreps are also known as *highest weights* in representation theory. The irreps of SU(2) are indexed by the half-integer $l \in \mathbb{N}/2$ known as *spin* in physics. We will denote these $(2l+1)$-dimensional irreps by R^l. Only the integer-spin R^l's descend to irreps of SO(3).

The finite-dimensional irreps of the Lorentz group, or more generally the real irreps of its double cover $\mathrm{SL}(2,\mathbb{C})$, are up to isomorphisms exactly the tensor products of representations of SU(2):

$$T^{(k,n)} = T^{(k,0)} \otimes T^{(0,n)} := R^{k/2} \otimes \bar{R}^{n/2},$$

where k, n are non-negative integers and the bar over $R^{n/2}$ indicates that this factor is acted upon by $\mathrm{SL}(2,\mathbb{C})$ via the conjugated representation (explicitly shown below). The dimensions of these spaces are $\dim T^{(k,n)} = (k+1)(n+1)$. The irreps of the Lorentz group are those $T^{(k,n)}$ for which $k+n$ is even.

Recall that the action of SU(2) on its spin l irrep is realized by the Wigner D-matrices $D^l(g)$, $g \in \mathrm{SU}(2)$. Due to the relation between $\mathrm{SL}(2,\mathbb{C})$ and SU(2), it is easy to parametrize the group elements using Euler angles. Introduce

$$\alpha = \varphi + i\kappa,\ \beta = \theta + i\epsilon,\ \gamma = \psi + i\varkappa,$$
$$\varphi \in [0, 2\pi),\ \theta \in [0, \pi],\ \psi \in [0, 2\pi),\ \kappa, \epsilon, \varkappa \in \mathbb{R}$$

(β and γ should not be confused with the velocity and boost factors from special relativity). These variables provide non-degenerate coordinates on $\mathrm{SL}(2,\mathbb{C})$, identifying it with the space $S^3 \times \mathbb{R}^3$. Any unimodular matrix $a \in \mathrm{SL}(2,\mathbb{C})$ can be factorized as

$$a(\alpha,\beta,\gamma) = \begin{pmatrix} e^{i\alpha/2} & 0 \\ 0 & e^{-i\alpha/2} \end{pmatrix} \begin{pmatrix} \cos\frac{\beta}{2} & i\sin\frac{\beta}{2} \\ i\sin\frac{\beta}{2} & \cos\frac{\beta}{2} \end{pmatrix} \begin{pmatrix} e^{i\gamma/2} & 0 \\ 0 & e^{-i\gamma/2} \end{pmatrix},$$

which is the complexification of the Euler factorization. Real angles parametrize the SU(2) subgroup, whereas the imaginary parts are essentially the rapidities parametrizing Lorentz boosts. This formula also expresses the so called

fundamental, or *defining*, representation of $\mathrm{SL}(2,\mathbb{C})$ acting on $T^{(1,0)} \cong \mathbb{C}^2$.

Furthermore, it is clear that the action of $\mathrm{SL}(2,\mathbb{C})$ on the irrep $T^{(k,0)}$ is given exactly by the analytically continued Wigner D-matrix of SU(2) spin $k/2$. Similarly, the action on $T^{(0,n)}$ is given by the conjugate representation of spin $n/2$. The conjugate representation of $\mathrm{SL}(2,\mathbb{C})$ of spin $1/2$ (the conjugate fundamental one) is given by $a \mapsto \epsilon\,\overline{a}\,\epsilon^{-1}$ where ϵ is the 2D Levi-Civita tensor. It is easy to check that

$$\epsilon\,\overline{a(\alpha,\beta,\gamma)}\,\epsilon^{-1} = \overline{a(-\alpha,\beta,-\gamma)}$$

(here the bar denotes complex conjugation). Combining these, we see that the action on $T^{(k,n)}$ corresponds to the tensor product of two Wigner D-matrices:

$$D^{k/2}(\alpha,\beta,\gamma) \otimes \overline{D^{n/2}(-\alpha,\beta,-\gamma)}.$$

For instance, on the fundamental representation of $\mathrm{SO}^+(1,3)$, for which $k=n=1$, these matrices are exactly the 4×4 Lorentz transformations (the defining representation of $\mathrm{SO}^+(1,3)$).

4. Principles of Equivariant Networks

Equivariant Universal Approximation Given two representations (V,ρ) and (V',ρ') of a group G, a map $F : V \to V'$ is called equivariant if it *intertwines* the two representations, that is:

$$F(\rho(g)\cdot v) = \rho'(g)\cdot F(v), \quad v \in V,\ g \in G.$$

Our goal is to design an architecture that can learn arbitrary equivariant maps between finite-dimensional representations of the Lorentz group. Even though the application described below requires only invariant outputs, the general way to achieve this is with an internally equivariant structure. First and foremost, this means having activations that are elements of linear representations of the group.

It was shown in (Yarotsky, 2018) that an arbitrary equivariant map between two completely reducible representations can be approximated by linear combinations of copies of a non-polynomial function σ applied to linear functions of G-invariants, with coefficients from a basis of G-equivariants (see Supplementary Material Section 4 for more details). Importantly, these polynomial invariants and equivariants are multiplicatively generated by a finite basis. This approximation theorem reduces our task to generating arbitrary polynomial invariants and equivariants for finite-dimensional representations of $\mathrm{SL}(2,\mathbb{C})$. In the Supplementary Material we show an extended version of the G-equivariant universal approximation theorem from (Yarotsky, 2018), which we paraphrase in simple terms here.

Theorem 4.1. *Given two completely reducible finite-dimensional representations V and U of a Lie group G, which can be* SU(2), SO(3), $\mathrm{SL}(2,\mathbb{C})$, *or* $\mathrm{SO}^+(1,3)$, *any equivariant map $\widetilde{f} : V \to U$ (including invariant maps for which $U \cong \mathbb{R}$) can be approximated by a feed-forward neural network with vector activations belonging to finite-dimensional representations of G that can iteratively perform the following operations:*

1. *Take tensor products of elements of representations of G;*
2. *Decompose tensor representations into isotypic components using the Clebsch-Gordan decomposition;*
3. *Apply equivariant linear maps between representations of G (as detailed in Section 4), including projections onto specific isotypic components;*
4. *Apply arbitrary sub-networks (such as multilayer perceptrons) to any G-invariants appearing after any of the above operations.*

Note that this theorem is a "Fourier space" statement (i.e. regarding networks based on irreps) extending the "real-space" characterization theorem proven in (Kondor & Trivedi, 2018).

Equivariant Linear Maps Now that we have established that tensor products are sufficient as the equivariant nonlinearity, we need to specify the form of equivariant learnable linear operations. Given a completely reducible representation V of G, we first find a linear isomorphism on V that represents it as a direct sum of its isotypic components: $V = \bigoplus_\alpha V_\alpha$ (the sum is taken over the labels α of all finite-dimensional irreps). Typically for us this isomorphism will be given by a Clebsch-Gordan operator. Each component V_α is itself isomorphic to a direct sum of zero or more copies of an irrep R_α: $V_\alpha = R_\alpha^{\oplus\tau_\alpha}$. We call τ_α the multiplicity of the irrep R_α in V. Now suppose the target representation can be similarly decomposed as $U = \bigoplus_\alpha R_\alpha^{\oplus\tau'_\alpha}$. Then, as was proven in (Kondor & Trivedi, 2018) by an extension of Schur's Lemma, all linear equivariant maps $W : V \to U$ are parametrized by a collection of $\tau'_\alpha \times \tau_\alpha$ matrices

$$W_\alpha \in \mathrm{Mat}(\tau'_\alpha, \tau_\alpha),$$

each of which acts on the list of irreducible components within an isotypic component V_α. This characterization (but not W_α's themselves) is independent of the choice of decompositions of the V_α's into irreducible components.

As demonstrated in (Cohen & Welling, 2016), the restriction to equivariant linear layers, compared to a general fully connected linear layer, leads to significantly fewer learnable parameters (depending on the representations at hand). Perhaps most importantly, *the loss function itself is G-invariant*. This means that if we transform the training samples $(\boldsymbol{x}_i, \boldsymbol{y}_i)$

by group elements $g_i \in G$, the trained weights W of an equivariant network will remain the same. In this sense, the weights are G-invariant, which makes them potentially interpretable as physical quantities.

Particle Interactions. As an elementary example of learnable Lorentz-invariant quantities in particle physics, the electron-muon scattering matrix element for initial and final 4-momenta p_1, p_3 of the electron, and initial and final 4-momenta p_2, p_4 of the muon, is given by

$$\mathcal{M}^2 \propto \frac{g_c^4 \left[p_1 \cdot p_3 + m_e^2\right] \left[p_2 \cdot p_4 + m_\mu^2\right]}{((p_1 - p_3)^2 - m_\gamma^2)^2}.$$

Here the dot products are taken with respect to the Minkowski metric, $m_e^2 = p_1^2$ and $m_\mu^2 = p_2^2$ are the masses, and g_c is an interaction strength parameter. Dot products are the invariant parts in the isotypic decompositions of tensor products of two 4-vectors, therefore a quantity of this kind can be very efficiently learned by an equivariant network if physically appropriate nonlinear activation functions are chosen. More complicated processes would involve higher nonlinearities like $(p_1^\mu p_2^\nu - p_1 \cdot p_2 \eta^{\mu\nu})^2$, which require several tensor products to be generated.

When a particle decay event produces hundreds of observed particles, generating all relevant Lorentz invariants (and even more so equivariants) up to a fixed polynomial degree quickly becomes an intimidating task that begs for a procedural solution. This is exactly the goal of our architecture.

5. Clebsch-Gordan product

The main nonlinearity in our equivariant architecture is the tensor product followed by a decomposition into irreducibles. This decomposition is known as the Clebsch-Gordan (CG) decomposition, and its coefficients in a certain canonical basis are called CG coefficients. We introduce the notation for the coefficients and a final formula for the CG coefficients of the Lorentz group here, but leave the details and derivations to the Supplementary Material. A reference for this material as regards SU(2) and the Lorentz group is (Gelfand et al., 1963).

Rotation group Let R_{l_1} and R_{l_2} be irreps of SU(2) of half-integer weights (spins) l_1 and l_2, respectively. Their product $R_{l_1} \otimes R_{l_2}$ decomposes via an isomorphism into a direct sum $\bigoplus_l \tilde{R}_l$, where $\tilde{R}_l$ are also copies of irreps of SU(2) and l ranges from $|l_1 - l_2|$ to $l_1 + l_2$ with unit step. This isomorphism is called the Clebsch-Gordan map

$$B: \quad \bigoplus_{l=|l_1-l_2|}^{l_1+l_2} \tilde{R}_l \to R_{l_1} \otimes R_{l_2}.$$

Since SU(2) is compact, its finite-dimensional representations can be assumed to be unitary with respect to the Euclidean norms on $\mathbb{C}^n$ (the resulting representation matrices are called Wigner D-matrices), therefore we can always choose B so that it is orthogonal.

For an arbitrary representation of SU(2) we define the *canonical basis* in it by $e_{l,m}$ where l ranges over the weights of the irreps contained in the representation, and for each l, the index m ranges over $-l, -l+1, \ldots, l$. Therefore the product space $R_{l_1} \otimes R_{l_2}$ has a basis induced from the respective canonical bases of the factors,

$$e_{l_1,m_1} \otimes e_{l_2,m_2}, \quad m_1 = -l_1, \ldots, l_1, \quad m_2 = -l_2, \ldots, l_2,$$

and the space $\bigoplus_l \tilde{R}_l$ naturally has the canonical basis

$$\tilde{e}_{l,m}, \quad l = |l_1 - l_2|, \ldots, l_1 + l_2, \quad m = -l, \ldots, l.$$

The CG coefficients $B_{l,m}^{l_1,m_1;l_2,m_2}$ are defined as the components of the CG map in these two bases:

$$B: \quad \tilde{e}_{l,m} \mapsto \sum_{m_1,m_2} B_{l,m}^{l_1,m_1;l_2,m_2} e_{l_1,m_1} \otimes e_{l_2,m_2}.$$

The summation is taken over all free indices occurring twice (and we will often omit mentioning them) over the ranges $|m_1| \leqslant l_1$, $|m_2| \leqslant l_2$, however $B_{l,m}^{l_1,m_1;l_2,m_2}$ vanishes whenever $m_1 + m_2 \neq m$ (see e.g. (Vilenkin & Klimyk, 1995, Ch. 4) for more on representation theory and CG coefficients of some classical groups).

Lorentz group The proper orthochronous Lorentz group $\mathrm{SO}(1,3)^+$ is isomorphic to the projective special complex linear group $\mathrm{PSL}(2,\mathbb{C})$. The Clebsch-Gordan map in this case is the isomorphism

$$H: \quad \bigoplus_{k,n} \tilde{T}^{(k,n)} \to T^{(k_1,n_1)} \otimes T^{(k_2,n_2)},$$

where the sum on the left is over

$$k = |k_1 - k_2|, |k_1 - k_2| + 2, \ldots, k_1 + k_2,$$
$$n = |n_1 - n_2|, |n_1 - n_2| + 2, \ldots, n_1 + n_2.$$

When an irrep $T^{(k,n)}$ of $\mathrm{SL}(2,\mathbb{C})$ is viewed as a representation of its subgroup SU(2), it decomposes into the direct sum of irreps (with unit multiplicities) $T^{(k,n)} \cong \bigoplus_{l=|k-n|/2}^{(k+n)/2} R_l$. This way, $T^{(k,n)}$ admits a *canonical basis*

$$e_{l,m}^{(k,n)}, \quad l = |k-n|/2, \ldots, (k+n)/2; \quad m = -l, \ldots, m.$$

In this basis, we define the CG coefficients for the Lorentz group by

$$H: \ \tilde{e}_{l,m}^{(k,n)} \mapsto \sum H_{(k,n),l,m}^{(k_1,n_1),l_1,m_1;(k_2,n_2),l_2,m_2} e_{l_1,m_1}^{(k_1,n_1)} \otimes e_{l_2,m_2}^{(k_2,n_2)}.$$

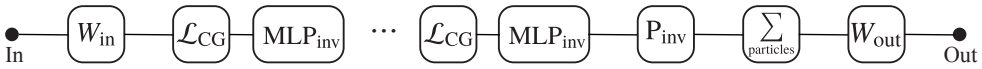

Figure 1. An elementary flow chart of LGN with Lorentz-invariant outputs. W_{in} is the linear input layer. It is followed by iterated CG layers $\mathcal{L}_{\text{CG}}$ defined in (2) alternated with perceptrons MLP_{inv} acting only on Lorentz invariants. The output layer projects onto invariants using P_{inv}, sums over particles for permutation invariance, and applies a linear layer. W_{in}, MLP_{inv} and P_{inv} act on each particle separately, but have the same values of parameters across all particles.

The CG coefficients can be expressed in terms of the well known coefficients for SU(2) introduced above:

$$H^{(k_1,n_1),l_1,m_1;(k_2,n_2),l_2,m_2}_{(k,n),l,m} = \sum_{m'_1,m'_2} B^{\frac{k}{2},m'_1+m'_2;\frac{n}{2},m-m'_1-m'_2}_{l,m} B^{\frac{k_1}{2},m'_1;\frac{k_2}{2},m'_2}_{\frac{k}{2},m'_1+m'_2} B^{\frac{n_1}{2},m_1-m'_1;\frac{n_2}{2},m_2-m'_2}_{\frac{n}{2},m-m'_1-m'_2} \times \times B^{\frac{k_1}{2},m'_1;\frac{n_1}{2},m_1-m'_1}_{l_1,m_1} B^{\frac{k_2}{2},m'_2;\frac{n_2}{2},m_2-m'_2}_{l_2,m_2}, \quad (1)$$

where the sum is taken over the following range of indices:

$$-\frac{k}{2} \leqslant m'_1 + m'_2 \leqslant \frac{k}{2}, \quad m - \frac{n}{2} \leqslant m'_1 + m'_2 \leqslant m + \frac{n}{2},$$
$$|m'_1| \leqslant \frac{k_1}{2}, \quad m_1 - \frac{n_1}{2} \leqslant m'_1 \leqslant \frac{n_1}{2} + m_1,$$
$$|m'_2| \leqslant \frac{k_2}{2}, \quad m_2 - \frac{n_2}{2} \leqslant m'_2 \leqslant \frac{n_2}{2} + m_2.$$

As always, the CG coefficients vanish when $m_1 + m_2 \neq m$. We provide more details on the derivation and computational implementation of this important formula in the Supplementary Material.

6. Equivariant Architecture (LGN)

We now describe the specific architecture that we applied to the problem outlined in Section 7. We call it the Lorentz Group Network (LGN).

Permutation Invariance Since the physics is independent of the labeling we put on the observed particles, the output of the network must also be invariant with respect to the permutations of the inputs. For our architecture this means that all learnable weights must be independent of the index of the input, and the simplest way to achieve it is with sums over that index at appropriate stages in the network. These sums are a key part of the architecture described here.

Input layer The inputs into the network are 4-momenta of N_{obj} particles from a collision event, and may include scalars associated with them (such as label, charge, spin, etc.). That is, the input is a set of vectors living in a $\left(T^{(0,0)}\right)^{\oplus\tau_0} \oplus T^{(1,1)}$ representation of the Lorentz group. Here, τ_0 is the number of input scalars. In this case, $\tau_0 = 2$ and the corresponding scalars are the mass of the particle and a label distinguishing observed decay products from the collider beams.

The input layer is simply a fully-connected linear layer acting on the inputs and producing $N^{(0)}_{\text{ch}}$ (number of "channels" at layer 0) vectors in each irreducible component. This layer acts on each input separately but the weights are shared between them to enforce permutation invariance.

CG Layers At the end of the input layer, we have N_{obj} activations $\mathcal{F}^{(0)}_i$, $i = 1, \ldots, N_{\text{obj}}$, living in $\left(T^{(0,0)} \oplus T^{(1,1)}\right)^{\oplus N^{(0)}_{\text{ch}}}$. We then apply a CG layer, iterated N_{CG} times, that performs tensor products, Clebsch-Gordan decompositions, and a learnable linear operation.

Assume that at the start of the p-th CG layer (starting with $p = 0$) we have N_{obj} activations $\mathcal{F}^{(p)}_i$ living in some representations of the Lorentz group (in fact our architecture guarantees that the representation is independent of i). The CG layer updates these activations $\mathcal{F}^{(p)}_i \mapsto \mathcal{F}^{(p+1)}_i$ according to the update rule

$$\mathcal{F}^{(p+1)}_i = \mathcal{L}_{\text{CG}}\left(\mathcal{F}^{(p)}\right)_i = W \cdot \left(\mathcal{F}^{(p)}_i \oplus \text{CG}\left[\mathcal{F}^{(p)}_i\right]^{\otimes 2} \oplus \oplus \text{CG}\left[\sum_j f(p^2_{ij}) p_{ij} \otimes \mathcal{F}^{(p)}_j\right]\right). \quad (2)$$

The Clebsch-Gordan operator CG follows every tensor product, and we are able to keep only the first few isotypic components to control memory usage. The last term models two-particle interactions via the pair-wise differences $p_{ij} = p_i - p_j$ while ensuring permutation invariance. The scalar coefficients $f(p^2_{ij})$ in this sum involve a function $f : \mathbb{R} \to \mathbb{R}$ with some learnable parameters, which weights the interactions of the i'th particle with other particles. The second term models a self-interaction of the i'th particle, and the first term simply stores the activation from the previous layer. W (also independent of i to ensure permutation invariance) is the equivariant learnable operator described earlier, and it mixes each isotypic component to a specified number $N^{(p+1)}_{\text{ch}}$ of channels. This choice controls the size of resulting vectors without breaking permutation invariance or Lorentz equivariance. To minimize computations, tensor products are performed channel-wise, which doesn't affect

expressive ability due to the presence of learnable linear operators mixing the channels.

MLP Layers Since Lorentz invariants can be freely transformed by arbitrary nonlinear functions without fear of breaking Lorentz symmetry, we can use traditional scalar neural networks each time any invariants are generated in our equivariant network. Namely, at the end of each CG layer we apply a multilayer perceptron to the $\left(T^{(0,0)}\right)^{\oplus N_{\text{ch}}^{(p)}}$ isotypic component. It takes $N_{\text{ch}}^{(p)}$ scalar inputs and produces the same number of outputs. The parameters of this perceptron are shared across all N_{obj} nodes in the CG layer. Adding these layers ensures that the layers of the network are non-polynomial.

Output Layer For permutation invariance, the output layer must take an arithmetic sum of the N_{obj} activations produced after the last CG layer. For a classification task, we are only interested in Lorentz-invariant outputs, therefore the output layer extracts the invariant isotypic component of this sum, and applies a final fully connected linear layer W_{out} to the $N_{\text{ch}}^{(N_{\text{CG}})}$ scalars, producing 2 scalar weights for binary classification:

$$\vec{w}_{out} = W_{\text{out}} \cdot \left(\sum_i \mathcal{F}_i^{(N_{\text{CG}})} \right)_{(0,0)},$$

where $()_{(0,0)}$ denotes a projection onto the spin-0 isotypic component (i.e. Lorentz invariants).

7. Experiments

We have tested the covariant LGN architecture on the problem of *top tagging*. This is a classification task that aims to identify top quark "jets" among a background of lighter quarks. Since the classification task is independent of the inertial frame of the observer, the outputs of the classifier should be Lorentz invariants.

Jets As explained in (Salam, 2010), high energy quarks produced in particle collisions lose energy through a cascading gluon emission process – a so-called *parton shower* – due to the structure of Quantum Chromodynamics (QCD), and eventually form stable hadrons that may be detected and measured. The lab frame in which those measurements are made may significantly differ from the parent quark's center-of-mass frame due to a Lorentz boost. In such a Lorentz-boosted lab frame, the parton shower process forms a collimated spray of energetic hadrons, depicted in 2, known as a *jet*. The jet 4-vector is related to that of its parent quark, as is the spatial and kinematic *structure* of the particles contained within the jet. A crucial task in collider physics is discerning the species of quark that has given rise to a particular jet. Approaches to this task involve the use of theory-inspired analytic observables, feed-forward neural networks, CNNs, recurrent neural networks, point clouds, and more. For a recent and comprehensive review, see (Larkoski et al., 2017).

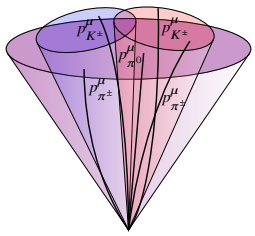

Figure 2. An example jet – as shown, different jet parameters (such as radius) may result in different clustering.

Dataset We perform top tagging classification experiments using the LGN architecture and the publicly available reference dataset (Kasieczka et al., 2019). This dataset contains 1.2M training entries, 400k validation entries and 400k testing entries. Each of these entries represents a single jet whose origin is either an energetic top quark, or a light quark or gluon. The events were produced with center-of-mass energy $\sqrt{s} = 14$ TeV, using the PYTHIA Monte Carlo event generator (SjÃűstrand et al., 2015). The ATLAS detector response was modeled with the DELPHES software package (de Favereau et al., 2014).

The jets in the reference dataset are clustered using the anti-k_t algorithm (Cacciari et al., 2008), with a radius of $R = 1$, where R is measured in (η, ϕ) coordinates. For each jet, the energy-momentum 4-vectors are saved in Cartesian coordinates (E, p_x, p_y, p_z) for up to 200 constituent particles selected by the highest transverse momentum $p_T = \sqrt{p_x^2 + p_y^2}$, where the colliding particle beams are aligned along the z-axis. Each jet contains an average of 50 particles, and events with less than 200 are zero-padded.

The 4-momenta in the dataset are all scaled by a uniform factor at the input to the network to avoid overflows and losses. As an extra pre-processing step, we add the proton beams to the list of particles as two 4-momenta of the form $(2, 0, 0, \pm 1)$ GeV[1]. The purpose of this is to fix an axis in each sample event, thus establishing a symmetry-breaking relationship between different samples. The energy (chosen to be 2 GeV) of these beams is somewhat arbitrary. Since these beams are distinct from the actual decay products in the dataset, we add a label to each particle, equal to $+1$ for the proton beams and -1 for all other particles. These labels are treated as Lorentz scalars.

[1]Special thanks to Jesse Thaler for this suggestion.

Hyperparameters For training, we performed a manual grid search. The main parameters are the number of CG layers, the highest irrep kept after each tensor product, and the numbers of channels. For top tagging, we found it sufficient to keep $T^{(k,n)}$ with $k, n \leqslant 2$, which means that the highest irrep is the 9-dimensional $T^{(2,2)}$ and the remaining irreps are $T^{(0,0)}$, $T^{(2,0)}$, $T^{(0,2)}$, and $T^{(1,1)}$. There were 3 CG layers, and the numbers of channels were chosen as $N_{\text{ch}}^{(0)} = 2$, $N_{\text{ch}}^{(1)} = 3$, $N_{\text{ch}}^{(2)} = 4$, $N_{\text{ch}}^{(3)} = 3$. The internals of the network are based on complex arithmetic, so these numbers should be doubled to count the number of real components.

The MLP layer after the p-th CG layer had 3 hidden layers of width $2N_{\text{ch}}^{(p)}$ each and used the "leaky ReLU" activation function. The scalar function f in 2 was a learnable linear combination of 10 basis "Lorentzian bell"-shaped curves $a+1/(1+c^2x^2)$ with learnable parameters a, b, c (each taking 10 values). The input 4-momenta were scaled by a factor of 0.005 to ensure that the mean values of the components of all activations would be order 1.

All weights were initialized from the standard Gaussian distribution. To ensure that activations stay of order one on average, the weights W were scaled down by a factor $N_{\text{ch}}^{(p)}/\tau_{(k,n)}$, where $\tau_{(k,n)}$ is the multiplicity of the $T^{(k,n)}$ irrep in the input to W. This ensures that W does not distort the values of the activations in higher irreps by orders of magnitude, making the contributions of various irreps unbalanced.

Performance and Cost The architecture was coded up using PyTorch and trained on two clusters with GeForce RTX 2080 GPU's. Each training session used one GPU and with the hyperparameters listed above it used about 3700MB of GPU memory with a mini-batch size of 8 samples. The wallclock time was about 7.5 hours per epoch, and our models were trained for 53 epochs each.

We compare the performance of our network to some of the other competitors (for a review see (Butter et al., 2019)). For each of these binary classifiers, we report four characteristics: the accuracy, the Area Under the Curve (AUC) score, the background rejection $1/\epsilon_B$ at the signal efficiency of $\epsilon_S = 0.3$ (ϵ_B, ϵ_S are also known as the false positive and the true positive rates, respectively), and the number of trainable parameters. Higher accuracy, AUC and $1/\epsilon_B$ are considered better. The mean and standard deviation in these metrics for LGN are reported based on 4 independent trained instances of the model.

The references for the algorithms listed here are: ParticleNet (Qu & Gouskos, 2020), P-CNN (CMS Collaboration, 2017), ResNeXt (Xie et al., 2017), EFP (Komiske et al., 2018), EFN and PFN (Komiske et al., 2019), TopoDNN (Pearkes et al., 2017). We should highlight EFP which constructs a special linear basis of polynomial observables that satisfy the so-called IRC-safety requirement in particle physics, and EFN which extends this idea to deep neural networks.

Table 1. Performance comparison between LGN and other top taggers that were measured in (Butter et al., 2019). Each performance metric is an average over an ensemble of networks, with the uncertainty given by the standard deviation.

Architecture	Accuracy	AUC	$1/\epsilon_B$	#Param
ParticleNet	0.938	0.985	1298 ± 46	498k
P-CNN	0.930	0.980	732 ± 24	348k
ResNeXt	0.936	0.984	1122 ± 47	1.46M
EFP	0.932	0.980	384	1k
EFN	0.927	0.979	633 ± 31	82k
PFN	0.932	0.982	891 ± 18	82k
TopoDNN	0.916	0.972	295 ± 5	59k
LGN	0.929 ± .001	0.964 ± 0.018	435 ± 95	4.5k

While our results do not match the state of the art, our model uses between $10-1000\times$ fewer parameters. More analysis of training and performance is provided in the Supplementary Material.

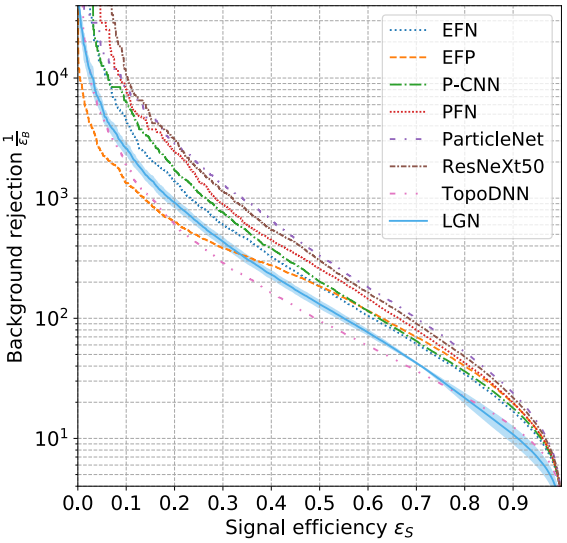

Figure 3. A comparison of an averaged ROC for LGN, against a sample of other top taggers. Higher is considered better. The ROC for LGN was sampled over the 4 trained instances of the model, with the error band width given by the standard deviation.

8. Conclusion

We have developed and successfully applied a Lorentz-equivariant architecture for a classification task in particle physics, top tagging. We believe this is the first application of a fully Fourier space equivariant architecture

in physics, following an chemistry application in (Anderson et al., 2019), and an important early step in building a family of physics-informed machine learning algorithms based on group theory. Symmetries have always been a central part of model-building in physics, and this work only further demonstrates the sharp need for symmetry- and geometry-based approaches to machine learning for scientific applications.

The performance of our neural network shines especially in terms of the number of learnable parameters. The trade-off is that an equivariant architecture takes more time to develop and its evaluation is more computationally intensive. However, once developed for a specific symmetry group, such as the Lorentz group or SL(2, $\mathbb{C}$) in our case, it is broadly applicable to many problems with the same symmetry at a very low development cost.

This network allows for many promising extensions in the context of particle physics. Future work will explore additional particle information such as charge and spin. The parameters of the model, which are Lorentz-invariant by construction, should be interpreted as physical quantities describing the particle decays. Permutation invariance can be further extended to permutation covariance. Another exciting problem is applying the network to regression tasks such as measuring masses of particles, or even 4-momenta. Finally, one could combine multiple symmetries such as the symmetry group of the Standard Model of physics (which includes U(1), SU(2) and SU(3)).

9. Acknowledgements

We thank Jesse Thaler for helpful advice on the training process. We acknowledge the significant support from the University of ChicagoâĂŹs Research Computing Center, Department of Computer Science, and especially the Center for Data and Computing (CDAC) for supporting this work through its Data Science Discovery Grant program. Finally, we thank the Flatiron Institute for hosting some of the authors during parts of the preparation of this paper. R. Kondor was supported by DARPA grant number HR00111890038.

References

Anderson, B. M., Hy, T., and Kondor, R. Cormorant: Covariant Molecular Neural Networks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *NeurIPS 2019*, pp. 14510–14519, 2019. URL http://papers.nips.cc/paper/9596-cormorant-covariant-molecular-neural-networks.

ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3(08): S08003–S08003, 8 2008. doi: 10.1088/1748-0221/3/08/s08003. URL https://doi.org/10.1088%2F1748-0221%2F3%2F08%2Fs08003.

ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett.*, B716:1–29, 2012. doi: 10.1016/j.physletb.2012.08.020.

Barut, A. O. and Rączka, R. *Theory of group representations and applications*. PWN—Polish Scientific Publishers, Warsaw, 1977. doi: 10.1142/0352.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. Spectral Networks and Locally Connected Networks on Graphs. In Bengio, Y. and LeCun, Y. (eds.), *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6203.

Butter, A., Kasieczka, G., Plehn, T., and Russell, M. Deep-learned Top Tagging with a Lorentz Layer. *SciPost Phys.*, 5(3):028, 2018. doi: 10.21468/SciPostPhys.5.3.028.

Butter, A. et al. The Machine Learning Landscape of Top Taggers. *SciPost Phys.*, 7:014, 2019. doi: 10.21468/SciPostPhys.7.1.014.

Cacciari, M., Salam, G. P., and Soyez, G. The Anti-k(t) jet clustering algorithm. *JHEP*, 04:063, 2008. doi: 10.1088/1126-6708/2008/04/063.

CMS Collaboration. The CMS experiment at the CERN LHC. *JINST*, 3(08):S08004, 2008. URL http://stacks.iop.org/1748-0221/3/i=08/a=S08004.

CMS Collaboration. Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. *Phys. Lett.*, B716:30–61, 2012. doi: 10.1016/j.physletb.2012.08.021.

CMS Collaboration. Boosted jet identification using particle candidates and deep neural networks. CMS-DP-2017-049, 9 2017. URL https://cds.cern.ch/record/2295725.

Cohen, T. and Welling, M. Group Equivariant Convolutional Networks. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd ICML*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, New York, NY, USA, 6 2016. PMLR. URL http://proceedings.mlr.press/v48/cohenc16.html.

Cohen, T., Weiler, M., Kicanaoglu, B., and Welling, M. Gauge equivariant convolutional networks and the

icosahedral CNN. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1321–1330, Long Beach, CA, USA, 6 2019a. PMLR. URL http://proceedings.mlr.press/v97/cohen19d.html.

Cohen, T. S., Geiger, M., KÃűhler, J., and Welling, M. Spherical CNNs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hkbd5xZRb.

Cohen, T. S., Geiger, M., and Weiler, M. A General Theory of Equivariant CNNs on Homogeneous Spaces. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 9142–9153. Curran Associates, Inc., 2019b. URL http://papers.nips.cc/paper/9114-a-general-theory-of-equivariant-cnns-on-homogeneous-spaces.pdf.

de Favereau, J., Delaere, C., Demin, P., Giammanco, A., LemaÃőtre, V., Mertens, A., and Selvaggi, M. DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *JHEP*, 02:057, 2014. doi: 10.1007/JHEP02(2014)057.

Diaconis, P. *Group representations in probability and statistics*, volume 11 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1988. ISBN 0-940600-14-5.

Erdmann, M., Geiser, E., Rath, Y., and Rieger, M. Lorentz Boost Networks: Autonomous Physics-Inspired Feature Engineering. *JINST*, 14(06):P06006, 2019. doi: 10.1088/1748-0221/14/06/P06006.

Esteves, C., Allen-Blanchette, C., Makadia, A., and Daniilidis, K. Learning SO(3) Equivariant Representations with Spherical CNNs. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. (eds.), *Computer Vision - ECCV 2018 Proceedings, Part XIII*, volume 11217 of *Lecture Notes in Computer Science*, pp. 54–70. Springer, 2018. doi: 10.1007/978-3-030-01261-8_4. URL https://doi.org/10.1007/978-3-030-01261-8_4.

Frankel, T. *The geometry of physics*. Cambridge University Press, Cambridge, second edition, 2004. ISBN 0-521-83330-2; 0-521-53927-7. An introduction.

Gelfand, I. M., Minlos, R. A., and Shapiro, Z. Y. *Representations of the Rotation and Lorentz Groups and Their Applications*. Graduate Texts in Mathematics. Pergamon Press, 1963. ISBN 9780080100692.

Hall, B. *Lie groups, Lie algebras, and representations*, volume 222 of *Graduate Texts in Mathematics*. Springer, Cham, second edition, 2015. ISBN 978-3-319-13466-6. doi: 10.1007/978-3-319-13467-3. URL https://doi.org/10.1007/978-3-319-13467-3. An elementary introduction.

Henaff, M., Bruna, J., and LeCun, Y. Deep Convolutional Networks on Graph-Structured Data. *CoRR*, 2015. URL http://arxiv.org/abs/1506.05163.

Kasieczka, G., Plehn, T., Thompson, J., and Russel, M. Top quark tagging reference dataset, 2019. URL https://zenodo.org/record/2603256.

Komiske, P. T., Metodiev, E. M., and Thaler, J. Energy flow polynomials: A complete linear basis for jet substructure. *JHEP*, 04:013, 2018. doi: 10.1007/JHEP04(2018)013.

Komiske, P. T., Metodiev, E. M., and Thaler, J. Energy Flow Networks: Deep Sets for Particle Jets. *JHEP*, 01:121, 2019. doi: 10.1007/JHEP01(2019)121.

Kondor, R. N-body Networks: a Covariant Hierarchical Neural Network Architecture for Learning Atomic Potentials. *CoRR*, 2018. URL http://arxiv.org/abs/1803.01588.

Kondor, R. and Trivedi, S. On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2747–2755, Stockholm, Sweden, 7 2018. PMLR. URL http://proceedings.mlr.press/v80/kondor18a.html.

Kondor, R., Lin, Z., and Trivedi, S. Clebsch-Gordan Nets: a Fully Fourier Space Spherical Convolutional Neural Network. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *NeurIPS 2018*, pp. 10138–10147, 2018. URL http://papers.nips.cc/paper/8215-clebschgordan-nets-a-fully-fourier-space-spherical-convolutional-neural-network.

Larkoski, A. J., Moult, I., and Nachman, B. Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning. 2017. doi: 10.1016/j.physrep.2019.11.001.

Monti, F., Boscaini, D., Masci, J., Rodolà, E., Svoboda, J., and Bronstein, M. M. Geometric deep learning on graphs and manifolds using mixture model cnns. In *CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5425–5434. IEEE Computer Society, 2017. doi:

10.1109/CVPR.2017.576. URL https://doi.org/10.1109/CVPR.2017.576.

Pearkes, J., Fedorko, W., Lister, A., and Gay, C. Jet Constituents for Deep Neural Network Based Top Quark Tagging. 2017.

Qu, H. and Gouskos, L. Jet tagging via particle clouds. *Phys. Rev. D*, 101:056019, Mar 2020. doi: 10.1103/PhysRevD.101.056019. URL https://link.aps.org/doi/10.1103/PhysRevD.101.056019.

Salam, G. P. Towards Jetography. *Eur. Phys. J.*, C67:637–686, 2010. doi: 10.1140/epjc/s10052-010-1314-6.

SjÃűstrand, T., Ask, S., Christiansen, J. R., Corke, R., Desai, N., Ilten, P., Mrenna, S., Prestel, S., Rasmussen, C. O., and Skands, P. Z. An Introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015. doi: 10.1016/j.cpc.2015.01.024.

Thomas, N., Smidt, T., Kearnes, S. M., Yang, L., Li, L., Kohlhoff, K., and Riley, P. Tensor Field Networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds. *CoRR*, 2018. URL http://arxiv.org/abs/1802.08219.

Vilenkin, N. J. and Klimyk, A. U. *Representation of Lie groups and special functions*, volume 316 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995. ISBN 0-7923-3210-5. doi: 10.1007/978-94-017-2885-0. URL https://doi.org/10.1007/978-94-017-2885-0. Recent advances, Translated from the Russian manuscript by V. A. Groza and A. A. Groza.

Weiler, M., Geiger, M., Welling, M., Boomsma, W., and Cohen, T. 3D steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *NeurIPS 2018*, pp. 10402–10413, 2018. URL http://papers.nips.cc/paper/8239-3d-steerable-cnns-learning-rotationally-equivariant-features-in-volumetric-data.

Worrall, D. E., Garbin, S. J., Turmukhambetov, D., and Brostow, G. J. Harmonic networks: Deep translation and rotation equivariance. In *CVPR 2017*, pp. 7168–7177, Honolulu, HI, USA, 7 2017. doi: 10.1109/CVPR.2017.758. URL https://doi.org/10.1109/CVPR.2017.758.

Xie, S., Girshick, R. B., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *CVPR 2017*, pp. 5987–5995, Honolulu, HI, USA, 7 2017. doi: 10.1109/CVPR.2017.634. URL https://doi.org/10.1109/CVPR.2017.634.

Yarotsky, D. Universal approximations of invariant maps by neural networks. *CoRR*, 2018. URL http://arxiv.org/abs/1804.10306.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep Sets. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3391–3401. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6931-deep-sets.pdf.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision - ECCV 2014 Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pp. 818–833. Springer, 2014. doi: 10.1007/978-3-319-10590-1_53. URL https://doi.org/10.1007/978-3-319-10590-1_53.